Characterising Dataset Search – an Analysis of Search Logs and Data Requests

Emilia Kacprzak[a,b,*], Laura Koesten[a,b], Luis-Daniel Ibáñez[a], Tom Blount[a], Jeni Tennison[b], Elena Simperl[a]

[a] *University of Southampton, UK*
[b] *The Open Data Institute, UK*

Abstract

Large amounts of data are becoming increasingly available online. In order to benefit from it we need tools to retrieve the most relevant datasets that match ones data needs. Several vocabularies have been developed to describe datasets in order to increase their discoverability, but for data publishers is costly to cumbersome to annotate them using all, leading to the question of what properties are more important. In this work we contribute with a systematic study of the patterns and specific attributes that data consumers use to search for data and how it compares with general web search. We performed a query log analysis based on logs from four national open data portals and conducted a qualitative analysis of user data requests for requests issued to one of them. Search queries issued on data portals differ from those issued to web search engines in their length, topic, and structure. Based on our findings we hypothesise that portals search functionalities are currently used in an exploratory manner, rather than to retrieve a specific resource. In our study of data requests we found that geospatial and temporal attributes, as well as information on the required granularity of the data are the most common features. The findings of both analyses suggest that these features are of higher importance in dataset retrieval in contrast to general web search, suggesting that efforts of dataset publishers should focus on generating dataset descriptions including them.

Keywords: Dataset Search, Vertical Search, Search Logs

1. Introduction

Data has become the most important digital asset in the world and its availability on the web is increasing rapidly. A growing number of organisations, mostly in the public sector, have set up their own data portals to publish datasets related to their activities. Similar trends can be observed in a variety of sectors. In the public sector, through initiatives such as Open Government Data (e.g. US Open Data portal[1], UK Open Data portal[2] etc.), data can generate social impact, improve public services, and increase transparency [47]. Specialised vendors in commercial sectors such as finances and marketing, co-exist alongside data marketplaces that connect supply and demand (e.g. data.world[3], Microsoft DataMarket[4], etc.). In science, an increasing number of datasets get published together with scientific publications, as open access and reproducibility become mainstream across subjects and research communities (e.g. Mendeley Data [5], Elsevier DataSearch Platform [6]). A study by Cafarella et al. [7] estimated more than one billion sources of data on the web as of 2011, counting structured data extracted from web pages. In 2015 the Web Data Commons project extracted 233 million data tables from the Common Crawl [29]. The ability to generate business value from data analytics offers competitive advantage in virtually every industry worldwide [28].

Data is used in a variety of professional roles. Whether it is a data journalist writing an article that compares government transparency in different countries, an app developer trying to expand into new markets, a business analyst searching for evidence to substantiate their report, or a scientist replicating an experiment, the first and foremost step all these professionals have to take is to find, or retrieve the most rel-

*The corresponding author could be contacted via email address: emilia.kacprzak@theodi.org and is based in The Open Data Institute, 65 Clifton Street, London EC2A 4JE Tel: 020 3598 9395

[1] `http://data.gov`
[2] `http://data.gov.uk`
[3] `http://www.data.world`
[4] `http://www.datamarket.azure.com`
[5] `https://data.mendeley.com`
[6] `https://datasearch.elsevier.com`

evant datasets for their needs. In previous work, we found that data practitioners (with different professional backgrounds and skills) face various challenges finding the data they need [25], relying on general purpose search engines, or asking other people for recommendations, thus, motivating the research and development of dataset search engines.

In the context of the Semantic Web, some efforts have been made to use Linked Data tools and vocabularies to improve the discoverability of datasets on the web. Kunze and Auer [26] defined the problem of dataset retrieval as a specialisation of information retrieval. A list of relevant datasets is returned, focusing on the particular case of RDF datasets for which filtering and similarity metrics based on shared vocabulary usage. However, in an open data scenario, data formats and models are heterogeneous, and the cost of mapping and transforming to RDF is often too high for publishers. An alternative is to compute or manually fill metadata descriptions of non-RDF datasets using an agreed vocabulary, for example, the Data Catalogue Vocabulary (DCAT) was designed to facilitate interoperability between data catalogues published on the web. This includes descriptions of keywords, theme, frequency, spatial and temporal coverage. Numerous extensions to DCAT have been developed to include additional properties that are considered relevant by their designers, e.g. DCAT-AP (for public sector data), GEO-DCAT-AP (geospatial properties) or Data-ID (versioning, technical descriptions of datasets). However, to the best of our knowledge, there are no systematic studies from the point of view of data consumers about what properties are more important than others for effective search and discovery of datasets. This is important, as the generation of metadata needs to be done on a property by property basis, which also represents a cost for data publishers. Knowing what are the properties that they need to focus on to satisfy user information needs reduces the time and effort required for the publishing process. Furthermore, current open data portal solutions base their metadata search on indexing free text descriptions of datasets and applying document modelling and search techniques. We believe that our findings could better inform what kind of advanced search functionalities should be explored to support user information needs.

In a nutshell, our goal is to advance towards the understanding of the most important properties of a dataset description from the point of view of data consumers, by analysing how people search for data on current portals. In particular we aim at answering the following **research questions**: (*a*) What are the characteristics of queries for datasets in terms of their length, distribution, and structure? How this informs the decision of which properties should be prioritized in a description and how they are currently represented in data description vocabularies? (*b*) How does the search behaviour of users in dataset search differ in comparison to general web search? (*c*) How do people request data when they are allowed to formulate their information needs with no restrictions. What properties do they consider the most important and how are these used?

In previous work [21] we presented an initial analysis focusing on internal data search queries on a portal. Internal queries are queries issued by users through the search functionality available on a data portal. In this work we extended our initial analysis and added two other data sources: external search queries and data requests, for a total of three. First we present a more in-depth analysis of internal search queries submitted to governmental open data portals with international scope (data.gov.uk and Office for National Statistics from the UK, open.canada.ca from Canada and data.gov.au from Australia). Secondly we analyse external search queries, which are queries issued on general web search engines that resulted in the user landing on one of these data portals (for both UK based portals).

Search functionalities tend to fall short for information seeking tasks for datasets [25] and queries issued on portals are too short to provide the basis for an extensive log analysis. Therefore we also analyse data requests. data requests are unstructured descriptions of datasets made by citizens directly to open data portals, they can be collected through web forms (*e.g.* data.gov.uk or datos.madrid.es), email (*e.g.* danepubliczne.gov.pl), or regular/quarterly platform meeting inviting the community (emphe.g. open.wien.gv.at). In this paper, we analyse a sample of 200 requests (from a corpus of 1600) submitted to data.gov.uk via a web form. Each data request consists of several fields; the structure can be seen in Section 3 in Figure 3.

Summary of findings. When characterising users (on the UK based portals) we found that data search is mostly a work related activity - portals are mostly accessed from desktop computers on weekdays during working hours. Returning users had much longer session durations, which suggests they might benefit from additional, more advanced, search functionalities. The majority of users ended up on data portals via external search engines. A number of users issue keyword queries that indicate a specific data portal. This suggests that search functionalities on portals might not be considered sufficient, which was suggested by qualitative analysis of interviews in our previous work [25].

Queries on data portals are generally short, which

might be the result of a lack of trust that longer queries will return useful results. This assumption was supported in our previous study Koesten et al. [25]. The results suggested that issuing a query is often conceptualised as an activity aiming to narrow down relevant subsets of data that is available on a topic, rather then expecting a matching dataset directly in the result list as we are used to from web search. Queries contained geospatial and temporal information - in both cases relevant keywords are represented in different levels of granularity (month/years, cities/regions). Queries including indications of time were five time more frequent than in web search. Furthermore data format and file type were popular amongst the queries - in case of external queries a fifth of all queries contained such attributes.

The analysis of data requests revealed similar features. The most common features mentioned in the requests were temporal and geospatial information often with definitions of their expected granularity. These features tend to be complex and need to be taken into account when generating metadata for datasets that is utilised by search functionalities. The wrong granularity in terms of both location and time can easily result in the data not being usable for a task. Furthermore, more than one dataset can be equally relevant to a single information need. Requiring information for longer time spans can result in many equally relevant datasets, as each contains a portion of the desired time period. Our findings suggests that publishers should focus their efforts on generating spatio-temporal properties of their descriptions, motivating the development of search interfaces for appropriately filtering and joining by them.

The remainder of the paper is structured as follows: Section 2 gives an overview of background and related work in vertical search engines and in query log analysis. Section 3 describes the data and methods used in our study. Sections 4, 5 and 6 report on the results of the analysis, following the research questions outlined earlier. Section 7 then discusses the main findings of our study and their implications, while section 8 points to the limitations of our work. Section 9 concludes the paper and outlines directions for future work.

2. Background & Related Work

In this section we give an overview of work that characterises different search verticals and compare them against dataset search, emphasising studies that leveraged query log analysis to understand users of those verticals. We further describe existing approaches in dataset search and discuss open data portals and current metadata standards to understand current search functionalities for data on the web.

2.1. Web Search

General Web Search. Web search engines take advantage of the specific structure of the web and use known text mining techniques (e.g. tf-idf [51]) to support search. General web search has evolved and is using sophisticated techniques such as machine learning [1], question answering systems [27] or personalisation of results for each user [39]. However, general web search is not a good fit for dataset search for several reasons: In their study of tables embedded in a web pages Cafarella et al. [8] pointed out that structured data on the web, cannot easily be mapped to unstructured text approaches. Tables lack the incoming hyperlink anchor text that is utilised in general web search. Web search algorithms (e.g. PageRank-based) are not applicable to the same extent to table search, particularly as tables of widely varying quality can be found on a single web page. The same constraints are applicable in a dataset search scenario as data is not an in-page element surrounded by additional context, but a source on its own.

Vertical Search. Vertical search is search in which the subject of the search is a specific subset of online content, as opposed to general web search where the aim is to include all types of resources. The subject of a vertical can be a collection which is distinct based on topic, data type or context. The differences in the underlying information source for each specific vertical require targeted information retrieval practices. Each vertical search strategy focuses on how to best utilize the additional information and structure that is provided with each information source. There are multiple distinct verticals which have been subject of specific query log analyses, for example people search engines [48], email search [2, 32], research publication search [30, 50] or search over linked data [16]. For instance, for **email search** Ai et al. [2] noticed that users know the precise attributes of a resource they are looking for. The authors point out that one of the key differences to general web search is that a set of emails is a personal set unique for each user. On top of that, email search offers additional metadata (e.g. senders' email address, subject or timestamp) which can help both organising and searching through the results. **People search** [48, 15] is gaining more importance with portals such as LinkedIn or Facebook. In this vertical, relevant factors are the first and last name of the person [15]. Search can also depend on the relations of two people that could be expressed

through the same educational background, home town or common friends. People search is also of importance in enterprises. This is slightly different as e.g. phone number, email or the organisation employing the person can become relevant which alters the way queries are issued [15]. In search for **research publications** Li et al. [30] and Yu et al. [50] argue that general web search does not fully take advantage of the potential of the specific subset of resources - temporal information attached to each of the publications. Although algorithms such as PageRank and HITS calculate the relevance of each resource and include this information while ranking the relevance of resources to a user query; both favour older resources over newer ones. In the publication search vertical, the reputation of the resource, in addition to its content relevance, citation count and reputation of its authors and journals are the most influential parts. However, in this scenario it is important to take the publication time bias into account when generating search results for the query. In **Product search**, the motivation is the increasing product specificity and variations in consumer preferences for optimizing online marketplaces [10]. Implementing solutions that support synonyms and a variety of languages used in order to find a specific products is more important than in other verticals [46]. Users in product search scenarios consider various facets (e.g. price of an item) to be of importance, depending on the product they are searching for. General search cannot fulfil scenarios in which a user wants to filter by items with a lower price than a set boundary (a specific value) [45]. This characteristic differentiates product search as a vertical from general web search.

2.2. Dataset Search

In this work we are analysing search for datasets published on the web, specifically on Open Data portals. There is a large body of research targeting search and exploration of tables embedded within web pages. Cafarella et al. [8] propose an approach for searching through web pages that contain a table with structured data. They include scores measuring the coherency of a table; this ranking was called *SchemaRank*. This is a starting point for a web table search algorithm that takes advantage of the web structure.

In their work on Fusion Tables, a data management system developed by Google [13], Google creates a corresponding HTML page for each table that is then crawlable by general search engines in the same way a regular web pages. Fusion Tables also recognised the need for an internal search functionality that will allow to search only among tables that are managed by the system since, as shown in a study by Cafarella et al. [8], existing techniques for dataset retrieval are not applicable for datasets.

Dataset search and retrieval on the web is a relatively unexplored area compared to document search and retrieval. Kunze and Auer [26] define the problem of *dataset retrieval* as determining the most relevant datasets according to a user query. They restrict their scope to RDF datasets and propose a retrieval mechanism inspired by faceted search, where a dataset relevance is checked against a set of semantic filters. Our work aims at determining through the analysis of user queries, which semantic filters are the most important to implement and show.

Open Data Portals. Open data portals are a point of free access to governmental and institutional data for both commercial and non-commercial purposes through cataloguing common metadata, which allows search for data. In this work we analyse the search logs of four open data portals as well as requests for data that have been made to data portals.

Mitlöhner et al. [31] presented an analysis of the data that can be expected on such portals. The average open data CSV file contains 365 rows and 14 columns and most values are numerical [31]. When analysing the notation in headers they found 40% contained underscores; 33% consisted of single words; 17% multiple words and 9% were expressed in camel case. The form in which headers are written can influence machine readability and potential further work with a dataset (e.g. making it harder to transform such a file into a RDF data model) [11].

Many open governmental platforms are based on **CKAN**[7], an open source data management system, which provides tools for publishing, sharing, finding and using data. It is used by many national data publishing entities (used by e.g. the UK, USA, Canada, Australia and European Data Portals) which includes three of the portals analysed in this work. There are other open data platforms, such as Socrata (used by e.g. the Chicago[8] and New York City government open data portal[9] or OpenDataSoft[10] (used by e.g. Open Data America[11]). CKAN is built using Apache Solr[12], which uses Lucene to index the documents. In this scenario the documents are the datasets' metadata provided

[7] https://ckan.org/
[8] https://data.cityofchicago.org/
[9] https://data.ny.gov/
[10] https://www.opendatasoft.com
[11] https://opendataamerica.com/pages/home/
[12] http://lucene.apache.org/solr/

by the publishers. The search functionality in **Solr** is composed of two main operations: finding the documents that match the user query and ranking those documents. After the final set of matching documents has been found, an additional operation is necessary to calculate a relevance score for each of the matching documents. **Lucene**[13] maps metadata fields into an inverted index consisting of a list of terms and ids of documents (dataset metadata) in which the given term appear. Calculating the relevance of a dataset to a user query is performed using the *term frequency–inverse document frequency* (**TF-IDF**) algorithm. It calculates the weighting of a term through a composition of two statistical approaches. TF is the frequency with which a word appears in a datasets metadata, whereas IDF indicates the inverse proportion of the word's frequency in the all set of a datasets metadata. The basic idea of this solution is that the more frequently a word appears in a document, i.e., in this case the metadata description about a dataset, the more accurately it describes its content [38]. On the other hand, if the word appears in more documents, it becomes less representative for a single document and should be given less weight [51]. Each datasets metadata and query are represented as a vector in a vector space model; the similarity score between them is the result of calculating a cosine between the query vector and the dataset metadata vector. Approaches for indexing datasets should benefit from their structured information. Using term frequency based approaches might miss the fact that we write differently in natural language than we structure information in a spreadsheet. The main topic or the key concepts are likely to appear more often in natural language throughout the text, but in structured documents those concepts might be mentioned only once. In this work our aim is to investigate directions for how dataset search could be improved based on user information needs and to rethink existing approaches by questioning their applicability in dataset search as a unique search vertical. Our query log analysis is generalisable to any data portal that offers a keyword search box. For data requests, we just require the request to be in an open text format.

Metadata Vocabularies and Standards. Defining metadata features for dataset search is important in order to tailor search functionalities to the specific characteristics of dataset search. Several standardization efforts have been undertaken for defining metadata about datasets on the web, that could be used by search engines. One of them is **DCAT**[14] which is used by the CKAN platform. DCAT is an RDF vocabulary used to describe datasets in data catalogues. It enables better interoperability between different data catalogues. It can be used to describe structured data on the web. Over time numerous extensions of the DCAT vocabulary were proposed in order to facilitate different needs. They present a wide coverage of properties, describing datasets from various angles, for example DCAT-AP for public sector data with country specific extensions, and GEO-DCAT-AP for geospatial properties. In our work we want to understand which of those properties are most important from a user perspective when searching for data. Detecting which properties are of higher importance form users perspective, allows to focus the selection of metadata properties for describing a datasets to generate. **The CSV on the Web** Working Group developed a standard[15] for expressing useful metadata about tabular resources and CSV files specifically. Their goal is to provide a standardised way of ensuring consistency of data types and formats (e.g. uniqueness of values within a single column) for every file. Furthermore, **schema.org** [14] is a schema that can be used for describing structured data on the web. It is applicable to a variety of data formats and is used as mark-up describing structured content (e.g. tables within web pages) or as a metadata schema describing data using a defined list of attributes[16].

There are efforts for integrating different standards into homogeneous data structure for existing metadata schemas such as [33], in order to allow better accessibility and discoverability of the datasets. Existing metadata schemas might not be extensive enough to provide sufficient background for search and discovery processes. This work aims to explore the characteristics of dataset search and inform metadata design decisions by uncovering patterns in dataset search to understand which of the properties in existing metadata is most important, or if there is any property currently missing.

2.3. Query Log Analysis

Analysing query logs serves as a proxy to analyse the search behaviour of users [23, 22, 49] and can serve as a way to understand the users intent when interacting with a search functionality [9]. The first query log analysis on the web was published in 1999, for the Altavista search engine [40], and the approach has since

[13] `https://lucene.apache.org/core/`

[14] `https://www.w3.org/TR/vocab-dcat/`

[15] `https://www.w3.org/TR/tabular-data-primer/`

[16] `http://schema.org/Dataset`

been used to study many aspects of web search (see [20] for a survey).

Search patterns have unique characteristics in different search environments, for instance Ortiz-Cordova et al. [35] analyse patterns in search behaviours within two sets of logs: internal and external search logs. Those sets were collected for the ecommerce site *www.BuenaMusica.com*, listing traffic coming from general search engines (external logs) and search activity within internal search function (internal logs). As shown in [18], in their transaction logs analyses of nine search engines, the results of different search log analysis are not directly comparable, however they can provide valuable insights into search behaviour. Several metrics for analysing search logs were developed for general web search; a summary can be seen in Table 1. Their relation to our analysis is detailed below.

Query Length and Distribution. These are the most commonly presented statistics also analysed in this work. Taghavi et al. [43] have shown two trends in web search query length and its distribution: an increase in query length over time and that the distributions of terms follows a power-law or Zipf distribution. Ortiz-Cordova et al. [35] show the difference in average length and length distribution between internal and external queries. The results show that internal queries are shorter than external queries (on average 2.76 words for external and 2.25 words for internal). They used this information further to analyse the differences in consecutive search activities and to define search patterns.

Query Types Classification. Broder et al. created a taxonomy of web search queries based on user needs [5]. We believe this taxonomy is not directly applicable to dataset search as the information need is *finding data* and could so be seen as predominantly informational. We chose to classify queries containing specific types of information that have been studied in other search contexts: acronyms, geographic location, temporal indication and numeric values. Understanding the number of queries related to these dimensions can help shape indexing strategies for dataset search engines.

User and Session Statistics. Information retrieval studies for web search commonly also analyse behavioural characteristics, which are not directly concerned with the query itself but with the users search behaviour. These can give additional insights about the user population who perform dataset search that cannot be obtained by analysing the query itself.

Query Structure. We found a negligible use of special operators in our dataset, so we did not conduct further analysis on this aspect. The composite and non-composite characteristics studied in [4] are for long queries. As one of our results shows, *dataset search* queries are typically short, we did not find this analysis suitable for our case. We did analyse the number of question queries [4] to find if users are asking questions in their queries or are merely searching for datasets.

Topics. The topic categorisation used in other studies [3] gives an overview of topics asked in web search queries. This is not directly applicable to our case, as data portals have a more limited scope. We analysed topics based on the categorisation proposed on the data.gov.uk portal (which is detailed in Section 3 Table 3).

3. Methodology

In our experiments we used three types of data acquired from different governmental open data portals (described in more detail in Section 3.1): internal and external search logs as well as data requests. We analysed how people ask for datasets using two complementary methods [6]: a query log analysis over search log data collected via Google Analytics and a qualitative thematic analysis of concepts represented in the data requests.

3.1. Search Logs

The logs primarily represent search for structured data. All portals collect log data using Google Analytics, but might be using different settings. As a consequence, and as they started recording their logs at different points in time the available information and time frames per portal vary. Four well-known data portals from three English-speaking countries - United Kingdom, Canada and Australia - provided their query logs to us: the official UK government Open Data portal (DGU), the Office for National Statistics of the UK (ONS)[17], the Australian government Open Data portal (AUS)[18] and the Canadian government Open Data portal (CAN)[19]. Three of the portals store datasets or links to datasets and use the CKAN portal software[20] (i.e.

[17] https://www.ons.gov.uk
[18] https://www.data.gov.au
[19] https://www.open.canada.ca
[20] http://ckan.org

Metrics used
Query Length & Distribution Average length, distribution, percentage of 1, 2 and 3 words queries [4, 43]
Query Structure Types of queries: *question*, *operator*, *composite* and *non-composite* [4]. Question queries: starting with words that indicate questions, e.g. *what* or *how*; operators:containing boolean operators like *AND*, *OR* and *NOT* or special web search operators e.g. *url*, *site* or *filetype*; composite: queries that could be represented as compositions of short queries; non-composite: queries that cannot be represented by short queries - these can be divided in *noun phrases* and *verb phrases*.
Topics Several topic classifications in literature [3, 41]. E.g.: Spinks el al. classification: *Commerce, travel, employment or economy*, *People, places or things*, *Unknown*, *Computers or Internet*, *Sex or pornography*, *Health or sciences*, *Entertainment or recreation*, *Education or humanities*, *Society, culture, ethnicity or religion*, *Government*, *Performing or fine arts* [41].
Query Types Classification Broders et al. query taxonomy based on user needs [5]. 3 classes: *informational* - get information about something, *navigational* - to reach a particular site and *transactional* - to perform a transaction.
User and Session Statistics Session length and search exits (further detail in section 4) [17]; browser statistics, server usage, distribution of queries for a time frame (e.g. per day, week or month) [43].

Table 1: Metrics from web search studies used in this study

DGU, AUS and CAN), which bases its search functionality on the Solr search platform[21]. Search functionality is provided through a search box in the portal. Queries are evaluated against textual descriptions and metadata text fields associated with the datasets. The ONS stores their data on a custom portal which is more targeted at presenting an analysis of the collected data, partially through visualisations. Both DGU and ONS present 10 results for a query per page. AUS and CAN show 20 datasets as results per page. All portals provide facets by which users can filter and browse the results. The summary of collected logs with dataset logs size and time frames for collection can be seen in Table 2. We had three types of information objects for analysis: *queries*, *sessions* and *users*. We distinguish between two types of **queries**: *internal* - queries issued directly in the search box of a portal; and *external* - web search queries that led the user to open a page of the data portal. External queries were only provided for DGU and ONS. A query object comprises the following fields: *search terms* and *total unique searches*. The search terms of a query are made of the string, i.e., the sequence of search keywords, typed into the search box (of the portal, for internal queries; and of the web search engine, for external ones). Data on **sessions** included device, browser, number of pages viewed, and session duration. A session is defined as a group of interactions within a given time frame that ends after 30 minutes of inactivity or at midnight. The session duration is computed as the difference between the time the user entered the portal and the time they entered the last page they visited before leaving the site. Concerning **users**, we also had access to statistics on new and returning users. As per Google Analytics' method, a user is detected by setting a cookie in a device or browser, therefore, it is an upper bound on the actual number of users. The portals described in this analysis did not have any additional event tracking, e.g., click-through data, configured.

[21] http://lucene.apache.org/solr/

Pre-processing. Search log data from both internal and external queries was pre-processed as follows:

The N-Gram Fingerprint method was used to clean the data as it can detect basic spelling mistakes which could be a swap of two or more letters within a word. For example a 2-gram string for the word *london* would be *do,lo,nd,on* and for 1-gram *d,l,n*.

Step 2 Discard outliers in terms of length. 99.9% of all queries had less than 19 words. Based on manual inspection, we considered longer queries to be likely the result of accidental pasting of text into the search box and discarded them from our analysis.

Step 3 Finally, we removed those external queries which were registered, but not specified. They were of two types: *(not provided)* and *(not set)*, according to the eponymous Google Analytics flag. The first are not specified due to the privacy policy of Google Analytics, while the second refers to traffic that did not occur as a result of a search, but via referral sites, direct links, or other search channels such as Google Maps and Google Images.

Internal Search Logs. Queries issued directly to the internal search capacity of a data portal into the search box. We have a total number of $2,245,574$ internal queries per portal excluding queries removed in Step 2, $724,095$ unique queries determined via clustering in Step 1 (data cleaning). The breakdown of internal queries per portal after pre-processing steps can be seen in Table 2.

External Search Logs. Queries issued through a general web search engines search as that lead to a page of the data portal. This set consisted of $1,101,201$ external

Portal	Internal		External				Time Ranges
	All	Unique	All	Unique	Not Set	Not Provided	
DGU	1,058,197	332,823	1,062,937	419,750	3,159	3,902,006	30/01/2013 - 31/08/2016
ONS	950,593	342,054	38,264	13,887	824	326,596	28/02/2016 - 31/08/2016
CAN	231,473	46,661	-	-	-	-	23/08/2015 - 23/08/2016
AUS	5,311	2,557	-	-	-	-	01/08/2016 - 31/08/2016

Table 2: Summary of search log data. For internal queries, column *all* refers to the total number of internal queries per portal excluding queries eliminated in Step 2, while column *unique* refers to the number of unique queries determined via clustering in Step 1. For external queries, column *all* shows the number of queries obtained after removing overlengthy queries (Step 2) and not provided and not set ones (Step 3). Columns *not set* and *not provided* show the number of not set and not provided queries. The column *unique* was calculated like for internal queries (Step 1)

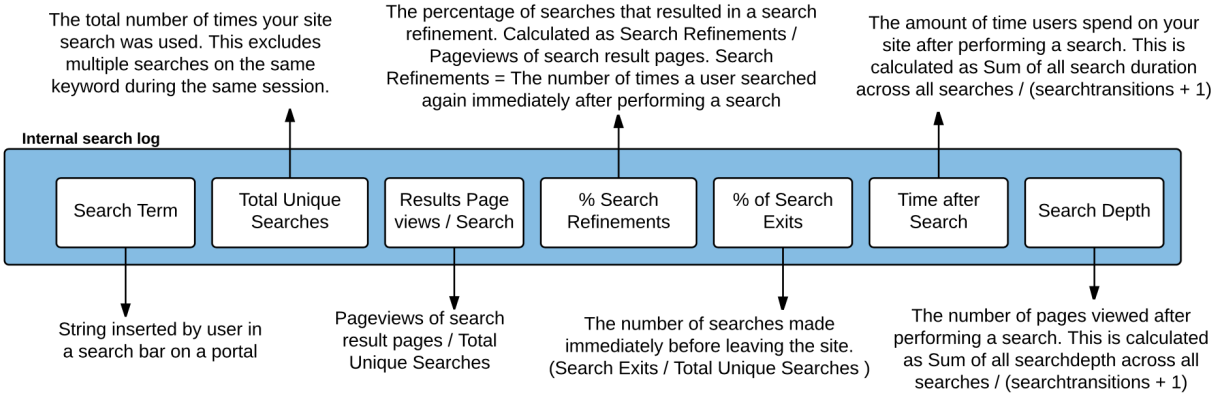

Figure 1: Structure of internal query logs and details of their meaning

queries, after removing lengthy queries (Step 2) as well as missing values from the data (Step 3). 3,983 were *not set* and 4,228,602 were *not provided* in our sample. There were 433,637 unique external queries (determined via data cleaning in Step 1). The breakdown of external queries per portal after pre-processing steps can be seen in Table 2. We assume that when an user issues a query to the search on an external engine, and clicked in the result that direct them to the data portal, their intention was to find a dataset. We acknowledge the limitation that we cannot know if the user just clicked on the dataset as part of an informational query looking for something else from the external portal. A possible more fine-grained heuristic is to collect and analyse the search exits from external queries, and assume that those that immediately exit the portal were not looking for a dataset.

3.2. Quantitative Analysis

This section describes the metrics presented in Table 3 chosen to analyse the queries. Their selection was based on background literature in web search and other search verticals shown in Section 2.3 and the relevance of commonly used metrics when applied to the analysis of dataset search logs. To analyse the data we first inserted all search logs into a MongoDB[22] database as separate entities. We created separate collections for internal search logs, external query log and data requests. Results for the metrics listed in Table 3 were generated using Python code[23] connected to the aforementioned collections unless specified differently. Results on user statistics were collected by using Google Analytics.

3.3. Data Requests

Data requests are a representation of information needs submitted by users of a data portal in order to get a specific dataset that they usually could not find. We used a set of 1600 data requests from the United Kingdom governmental open data portal (DGU) which are partially available as a dataset on the portal[25]. Requests represent information needs for data. As they

[22] http://www.mongodb.org
[23] `https://github.com/chabrowa/search-log-analysis`
[24] https://www.surveymonkey.com/mp/sample-size-calculator/
[25] `https://data.gov.uk/dataset/data-requests-at-data-gov-uk`

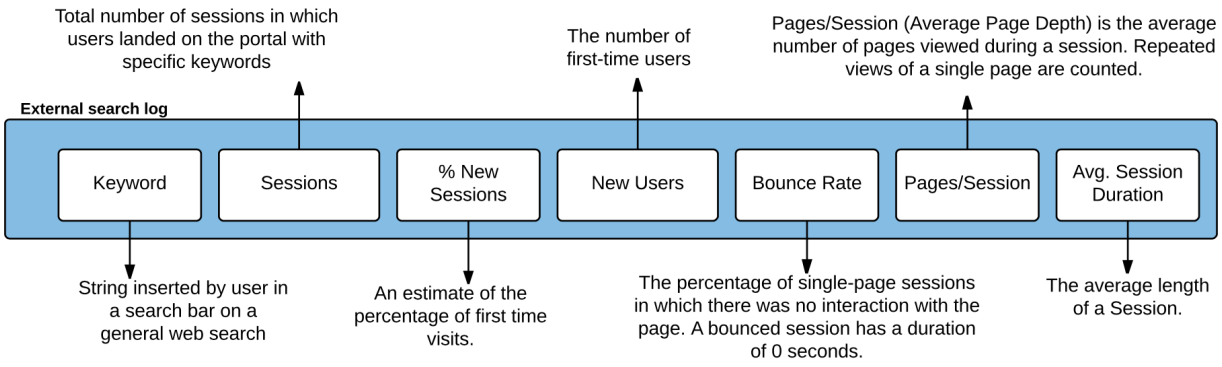

Figure 2: Structure of external query logs and details of their meaning

are submitted in natural language and are much longer than search queries they contain additional information of the task that people aim to do with the requested data. After contacting the portal owners we were able to acquire 800 additional requests that were tagged to remain confidential, however only publicly available elements of the data requests are included in the examples cited in this work.

Structure. Data requests are submitted to the portal via a semi-structured contact form available on the portal website by users who aim to satisfy their information needs. The form is targeted at citizens who can ask there for particular government datasets to be made public, and if possible, to have an open licence attached to it. Users are required to provide a description of the data along with a reason and context for their requests. This results in long descriptions of the data that is needed - providing us with additional context of the connected task. A sample description of the data request is for instance: *I wish to find out up to date figures of numbers of adults with moderate, severe and profound disabilities (in particular learning disabilities) who are currently working in the UK, either part time or full time; where in the UK they work and in what numbers ; and at what occupations*. We present an overview of the structure of all fields in a request form in Figure 3.

Pre-processing. Data requests in our sample were selected manually, we excluded requests that do not define a clear data need or that require complex data analysis. In addition, we filtered out the same request that were accidentally submitted twice as a separate requests to the portal. Our analysis was conducted over a set of 200 data requests which were randomly selected and met our inclusion criteria.

3.4. Qualitative Analysis

Thematic analysis. We analysed 200 data requests qualitatively, using thematic analysis - a method to identify patterns or themes within qualitative data [37]. Coding was done using NVivo (version 11), a qualitative data analysis package. Two of the authors individually coded a sample of the data requests inductively [44], compared code lists with each other and discussed conflicting results with two senior researchers. This process was repeated twice until there were no conflicting codes between the two researchers and there were no new emerging codes. These were then used to code the remaining data requests. We grouped emerging codes related to the *attributes* of the data and those related to the *structure* of the request into these two high level categories. For each of the categories we applied two layers of coding [37]. The *data attributes* layer allows an understanding of how users are talking about data when describing information needs to another person (the receiver of the data request is an employee of data portal). These are grouped into subsets of: geospatial content, temporal content, restrictions on the requested data (for instance specific formats or licences), mentions of the required granularity. The *request context* layer includes the prevalence of common features to get an overview of the composition of the data requests. This includes mentions of expected representation and structure of the data, the unit of interest (whether a data point, a dataset or the result of an analysis is requested), rationale for the data request or mentions of quality issues with existing datasets with a request for the same data, but in better quality.

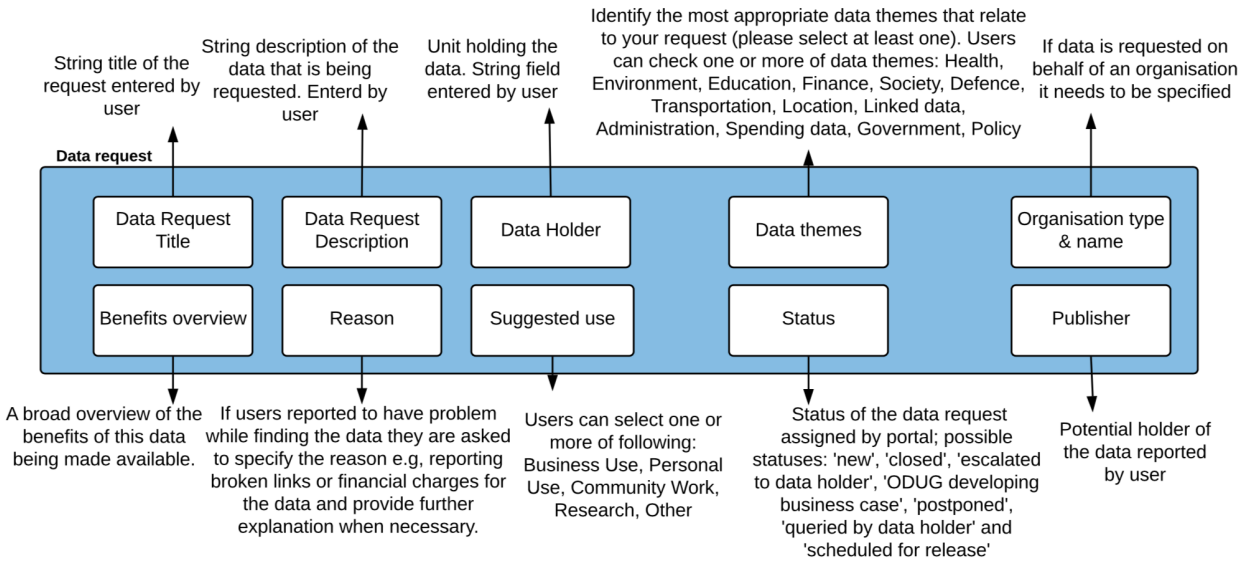

Figure 3: Data request fields and their descriptions; Description obtained from the data request form on DGU. This form is no longer used. The difference between 'data holder' and 'publisher' is unclear, both fields were used in the same way by the users of the system.

4. Users

In this section, we characterise the context of access of users to data portals. For this section we use session data from DGU and ONS.

Location. As expected from portals with a national scope, the majority of users (82%) were from the United Kingdom. 26% were from London, while other major cities, such as Manchester or Birmingham represented around 1 to 2%, with the rest of the access evenly spread across the country.

Devices. Table 4 shows how people access the portals. We distinguish between *desktop computers*, *mobile devices*, and *tablets* and list the relevant share of sessions for each of them, as well as the average number of pages viewed per session and the average session duration. An overwhelming majority are desktop computers (85% on average for both portals), compared to mobile devices ($\sim$ 8%) and tablets ($\sim$ 6%). Both the number of pages viewed and session duration are highest for desktops. We believe the high percentage of desktop users can be explained by the fact that data search is mostly a work activity; this is confirmed by the the time of the day people access the portal (see below). Looking for datasets is a first step in a much more complex workflow, in which a relevant dataset is subsequently downloaded and then inspected and visualised using exploratory data analysis tools. These activities are typically performed on desktop computers due to their larger screens and additional processing power.

Device	% Sessions		Pages viewed		Session duration	
	DGU	ONS	DGU	ONS	DGU	ONS
Desktop	79.81	90.95	3.42	3.04	02:35	03:41
Mobile	12.93	4.94	1.78	2.65	00:57	02:14
Tablet	7.27	4.11	2.24	2.29	01:22	01:53

Table 4: Devices used to access data portals

Time of access. Users are mostly active during weekdays, as can be seen in Figure 4. Monday has the highest level of activity, which falls slightly every day until Friday, to reach the lowest point on Saturday and grow slightly again on Sunday. Activity during weekends is approximately half or a third of that during week days. Users access the portals during working hours (8am to 6pm, local time) and issue most queries between 9am to 11am. This pattern, in combination with the prevalence of desktop computers, further suggests that dataset search is a work time activity.

Channels. As can be seen in Table 5, the majority of users (62.32% for DGU; 74.33% for ONS) reach both portals through the result page of a web search engine (a scenario which we refer to as *external*); by accessing the portal directly through its URL (*direct* - 14.3%

Metric - Method
User statistics Includes information about: devices, browsers, channels through which user reach the portal, user location, time of accessing the portal, ratio of new to returning users. Statistics gathered by Google Analytics
Search exits Number of sessions in which the user leaves the page immediately after searching through the portals search box, without clicking on any of the results.
Time after search The average amount of time users spent on the portal after performing a search, it is calculated as the sum of all search durations (including refinements) divided by the number of search sessions.
Search refinement Number of searches performed following an initial search within the same session, different from the initial query.
Average length; number of words in a query Computed for all internal and external queries. Both of these metrics were calculated for *all* queries in the log as well as for the subset of *unique* queries.
Query characteristics Matching queries to keywords describing: location; time frame; file and dataset type; numbers; abbreviations. Keywords used for each of those metrics are specified in Table 9. Computed for internal and external queries. The keywords for each category were selected by taking a sample of top 50% of queries and listing the words indicating particular information type (as listed in Table 9); we in addition used the most popular words that were not found in those top queries (e.g. yearly or quarterly). We compared the list of keywords against all queries to detect how many of them contained particular keyword.
Question queries To recognise question queries we counted queries containing the words: *what*, *who*, *where*, *when*, *why*, *how*, *which*, *whom*, *whose*, *whether*, *did*, *do*, *does*, *am*, *are*, *is*, *will*, *have*, *has* as done in [4]. Computed for all internal and external queries.
Query topical distribution Manual categorisation of topics was done by two of the authors for a sample set, representing the 665 most popular queries. This sample size was determined using a 99% confidence level, a 5% confidence interval (or margin of error - e = 0.05), z-score equal 2.58 (used for a 99% confidence level), distribution 50% (p = 0.5), which gives the largest sample size, and population size of 2.2 million queries using the following formula[24]: $$samplesize = \frac{\frac{z^2 * p(1-p)}{e^2}}{1 + \frac{z^2 * p(1-p))}{e^2 N}}$$ We derived 12 topics (plus *other*) from themes used by DGU to tag datasets. We exclusively categorised each query to one of these topics: *Business and Economy*, *Environment*, *Mapping*, *Crime and Justice*, *Government*, *Society*, *Defence*, *Spending*, *Towns and cities*, *Education*, *Health*, *Transport* and *Other*.

Table 3: List of metrics performed in the qualitative analysis

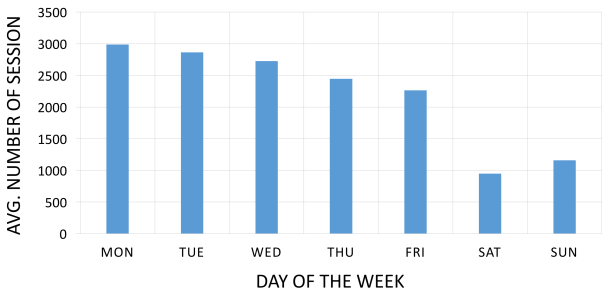

Figure 4: Distribution of sessions with search per weekday

for DGU; 16.72% for ONS); or by following a link from a different website that is not a social network or a search engine (*referral* - 2.62% for DGU; 8.52% for ONS). Less than 1% of visits are generated through social networks. The *Other* row in the table groups together traffic coming from email links, advertising, and paid search. The high share of externally driven traffic suggests that most users either resort to common web search engines to search for data and are then directed to the portals, or they use web search engines as proxies - this means, instead of going directly to data.gov.uk and issuing a search there, they start with a regular web search engine and add additional keywords to their queries, for example "data UK" which lead them to a portal. We discuss this type of query in the following section.

Browsers. The majority of data search sessions used Chrome (41.35%), followed by Internet Explorer (IE) (30.50%), Safari (13.97%), and Firefox (9%). Compared to general web browser usage,[26] both worldwide and from the UK, we note a higher share of IE users by almost 10%. As discussed earlier, people seem to be accessing data portals during weekdays and during office hours. In corporate and government environments, the use of IE is still widespread, which might help explain its relatively high popularity in the search logs.

New and returning users. Table 7 shows the the percentage of new and returning users and compares the two cases in terms of the average number of pages viewed and average session duration. Our main observation is that returning users view on average more pages and engage in longer sessions. Query log analysis from other verticals do not consider this metric except for [48], which reports 7% returning users out of its 7 million sessions. We believe these differences sug-

[26] Using statistics from 2013 to 2015, from `http://gs.statcounter.com/`

Channel	% Sessions	
	DGU	ONS
External	62.32	74.33
Direct	14.30	16.72
Referral	9.62	8.52
Social	0.86	0.43
Other	4.30	<0.01

Table 5: Channels through which users access portals

Browser	% Sessions	
	DGU	ONS
Chrome	37.95	44.76
Internet Explorer	29.87	31.13
Safari	16.09	11.86
Firefox	10.93	7.18
Other	5.16	5.07

Table 6: Browsers used to access portals

gest that users return with the intent to work with data and spend more time in assessing the relevance of their search results. The higher proportion of returning users to the ONS portal is probably a function of the reputation of the ONS as an established, authoritative source of data, compared to the much newer initiative around data.gov.uk. This was also confirmed by interviewees in [25], who said that they trust the ONS to deliver high-quality data that is useful in various scenarios. This trust is probably the case also for new users of the ONS portal, who on average spend more time on this portal than new DGU users.

User	% Sessions		Pages viewed		Session duration	
	DGU	ONS	DGU	ONS	DGU	ONS
New	76.47	58.38	2.63	2.92	01:38	02:50
Returning	23.53	41.62	4.54	3.1	04:08	04:32

Table 7: Percentage of new versus returning users per portal

Search exits and refinements. More than a fifth (21.34%) of the searches on DGU resulted in a search exit; for ONS this figure was 51.72%. A report on the UK government website, which covers the entire online presence of the government[27], reports the share of search exits at 11% and search refinements at 30%, based on logs collected over one month. In dataset search, these metrics are much higher, which could mean that the users did not find what they were looking for, and left the portal as a result. A search refinement was recorded in 22.77% of the sessions for DGU and 36.08% for ONS. We did not have access to the subsequent refined queries, which would have helped to shed light on the refinement strategies of the users and also the successfulness of their attempts.

5. Internal & External Queries

In this section we present the metrics introduced in Table 3 in section 3.2. We present the comparison of *internal* queries for four different open data portals and of *external* queries for the two UK portals for which they were available. Internal queries were the queries issued directly to the search functionality of the data portal whereas external queries are those issued to web search engines that lead users to open a page on the data portal. The assumption is that if a user opened a page in a data portal following a web search hit, the intention of the query was to retrieve a dataset.

Internal queries were analysed as one set and further details using specific measures are presented in this section. We categorised external queries in two categories. We refer to *proxy queries* when they contain the name of a data portal. All remaining external queries are referred to as *direct queries*. 6.71% of external queries for DGU and 54.82% for ONS are proxy queries. A proxy query indicates that the user wanted to reach a result from the portal in question, but did so through a web search engine instead of going first to the portal and use its search capability. Our initial analysis of proxy queries revealed a high variance of spelling and use of URIs. To avoid skewing results due to noise, we chose to focus on direct queries, excluding queries identified as proxy queries. We split the queries into direct and proxy queries by analysing keywords indicating portal names (i.e. queries containing word groups as: *gov* and *uk*; *office*, *national* and *stat* or *o n s*) or queries in a form of an URL link (i.e. queries containing *www* or *http*). Code used in order to split queries is available on Github[28].

Proxy queries were not considered in the analysis of external queries ; web-addresses, as well as spelling variations or incorrect typing, would result in a large amount of noise, which would skew the results. Therefore we only included *direct queries* in our analysis of external queries.

[27] `http://gdstechnology.blog.gov.uk/2014/12/22/monitoring-search-performance-on-gov-uk/`, accessed in January 2017.

[28] `https://github.com/chabrowa/search-log-analysis/blob/master/database/externalProxyQueries.py`

5.1. Query Length

Table 8 shows the average query length for both internal and external queries. Internal queries are between one and three words long, with an average of 2.03 words for all queries (median equal 2) and 2.67 for the unique ones (median equal 3). The average external query length is 3.98 words for all queries (median equal 4) and 4.74 for unique queries (median equal 4). External queries are on average more than one word longer than internal queries. This could be the result of web search queries being generally longer [43], or a different perception of the internal search functionality by users. External queries were found to be longer than the reported average of 3.08 words in a web search query by [43] in 2012. However, this might not fully apply currently as general web search underwent rapid developments, for instance in answering conversational search queries. These advances might have resulted in much longer queries[29].

Portal	Internal		External	
	All	Unique	All	Unique
DGU	2.04	2.78	4.12	4.82
ONS	2.52	3.42	3.83	4.66
AUS	1.63	2.31	-	-
CAN	1.93	2.17	-	-

Table 8: Average number of words per query

Figure 5 shows the distribution of internal queries according to their length, for all portals, for both all and unique internal queries. When considering all queries, single word queries represent almost half of the entire corpus. When focusing just on unique queries, this number falls to 25%. The distribution for unique queries is very similar to the results reported for web search engines in 2001 by Spink [42]. Figure 6 shows the distribution of external queries according to the number of words in a query, for ONS and DGU. The distribution of number of words in external queries is more similar to the one shown for general web search than the distribution of number of words in internal queries. It could be that advances in dataset search will lead to similar behaviour patterns as observed in web search today, which would mean longer queries, that are closer to natural language.

[29] https://searchengineland.com/google-hummingbird-172816

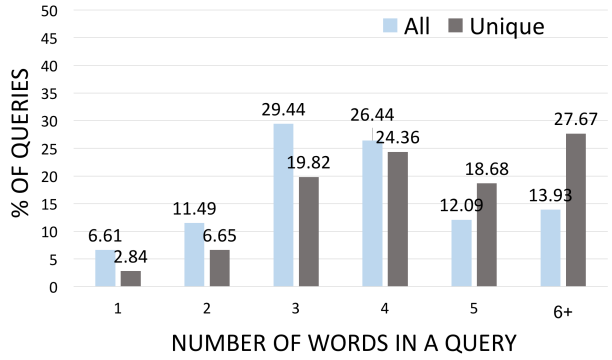

Figure 5: Percentage of internal queries by average number of words (all and unique queries, all portals)

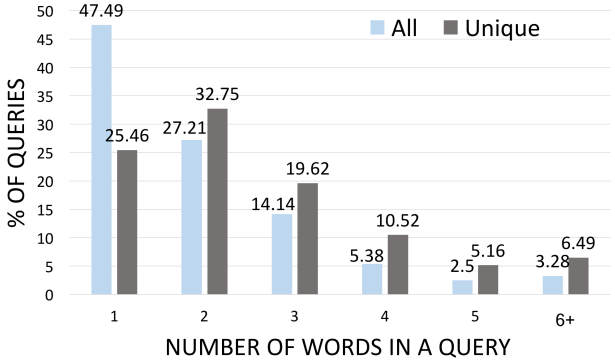

Figure 6: Percentage of external queries by average number of words (all and unique queries, ONS and DGU portals)

5.2. Query Types

In previous work we recognised that people have different strategies while searching for datasets [25]. In this work we focus our analysis on the keywords that are used in search queries for data. For most metrics as defined in Section 3, we computed the corresponding values automatically, for all internal queries of the four portals and all external queries for two UK based portals. To classify queries into specific topic we used a sample of the logs, as explained in Section 3.2 Table 3 under *query topical distribution*. Table 9 summarises the percentage of queries representing each metric.

Geospatial. In the set of internal queries we found 5.44% containing location specific keywords, whereas in the set of external queries this appeared to be slightly more popular with 7.93%. A previous study on general web search [12] reports this metric at 12.01%. This difference could be caused by the fact that the data portals

we studied are already bound to a country, and users do not need to specify the location as frequently as in general web search. In addition, it might be caused by the fact that queries in dataset search are significantly shorter in comparison to web search. With almost 50% of queries being one word queries, users might specify only the general topic of their information need in a query.

Temporal. Keywords indicating temporal information appeared in 7.29% of the internal queries and in almost twice as many external queries (12.26%). This number is much higher than the 1.5% reported for general web search [34]. This may mean that users have a higher interest in the temporal details related to the data they are searching as opposed to searching for web pages. This can include the time frame the data represents (data about a particular year) or the creation time of a dataset (the time the data was collected and published, including the frequency of updates).

File or data type. In dataset search, queries included restrictions for the shape the data should have, in order to fulfil the information need. 6.25% of internal and 20% of external queries included an indication of file type or data type. We note that the governmental portals represented in this study offer filtering options for file types that are not reflected in our data - the overall number of queries specifying a file format could be higher. However, it could also indicate that filters in current interface design might not be prominent enough. From a data point of view, this figure could be an indication that users search for alternative file types and formats and that publishers need to be able to support different and popular formats for their data. We believe the higher percentage in external queries is due to the fact that users intend to find data and not textual documents. In general web search, users need to indicate this in addition to their query - this step is unnecessary on data portals, which are designed to support search specifically for data.

Numerical. Data is often numerical - it was shown in [31] that numerical values are the most popular data type in open governmental datasets. In comparison to documents, which are mostly text, we also computed the number of queries containing numbers (excluding those indicating temporal information). 5.23% of internal queries contain any number and 0.38% contain only numbers. External queries present almost the same statistics with 5.23% for queries including numbers and next to zero for queries containing only numbers (only 0.008% of queries which we believe is so small due to the fact that external queries were much longer on average in comparison to internal queries). Those results show a disproportion in the amount of queries with numbers in comparison to the number of numerical values in the data.

This also indicates the need to understand the underlying meaning of numerical columns in datasets e.g. by lifting them to a linked data format which could then provide additional context to the data.

Abbreviations. In our analysis, we also identified that users frequently use abbreviations in their queries, as many datasets use acronyms like *rpi* for *Retail Price Index*. 5.11% of internal and 7.05% of external queries contained at least one acronym. However, we noticed that the full expansion of those acronyms is also used in queries. For the majority of governmental open data portals the main content that is indexed by the platform and searched over is a description of the dataset provided by the data publisher. Therefore some datasets are described in the index only with the full expansion of the acronym related to them. However, some users might only search using acronyms which results in false negative results.

Metric	internal	external
Geospatial - the name of a city or geographical area (either town, city, county, region or countries)	5.44%	7.93%
Temporal - years (1000 to 2017), names of months, days of a week and the words *week(ly)*, *year(ly)*, *month(ly)*, *day(ly)*, *date*, *time* and *decade*	7.29%	12.26%
File or data type - file types: *csv, pdf, xls, json, wfs, zip, html, api* and keywords denoting a type of dataset: *data, dataset, average, index, graph, table, database, indice, rate, stat*	6.25%	20.01%
Numbers - the number of queries including numbers excluding those indicating time frames	5.23%	4.46%
Only numbers - queries that contain only numbers	0.38%	0.008%
Abbreviations - 72 most popular, manually identified acronyms	5.11%	7.05%

Table 9: List of metrics with definitions and their prevalence in internal and external queries. The keywords lists are available on the Github repository.

Question queries. Question queries are increasingly more common in web search queries, thanks to advances in speech recognition and conversational search interfaces. This is not yet the case for dataset search -

less than 1% of internal queries in our logs are question queries. All figures are significantly below the 7.49% reported by [4] for web search. However, external question queries totalled 5.09% for DGU and 1.52% for ONS, significantly more than internal question queries and much closer to the results reported to general web search. We believe this is mostly due to users' understanding of the search functionality of dataset search engines (as a source of data to be downloaded for further use, and not as a question-answering engine) and due to the type of service that is currently provided, which might supply relevant search results for keyword queries but does not support question queries. (For instance the CKAN platform, as one of the most common data management systems, up-to-date does not provide such a functionality.)

Query topics. This metric aims to capture the domain of data people are searching for. We present our own classification, as the ones used in web search are not directly relevant [19] - for example, in web search sexual topics/pornography is the most prevalent topic category (25%) which does not fit the content of the platforms like governmental open data portals. In this work we used alternative topic categories as described in Section 3. Figure 7 shows the distribution of queries according to the data domain. The most popular category is *Business and Economy* (20.03%), followed by *Society* (14.74%). This is in line with our observations earlier about the use of data portals in professional contexts and is influenced by the nature of the portals themselves, which publish official statistics or data produced by different governmental departments. The distribution of topics for external queries differs from the one for internal queries. As can be seen in Figure 7 *Business and Economy*, *Environment* and *Other* queries are less frequent, while *Towns and Cities*, *Health*, *Education and Society* are more frequent. These are naturally influenced by the domain specificity of the portals.

6. Data Requests

We performed an in-depth thematic analysis, as described in Section 3.4, of the title and description of the data request. This is done on different levels, defined as *data attributes* and *request context* level, which are discussed in detail below. Below we present statistics over the different data request's multiple choice fields. These give us an overview of the people who issued the requests, of the most popular themes, and of the intention of use of the requested data to give context to the findings presented in Sections 6.1 and 6.2.

Organisation type	% of requests
Individual	45.43
Academic or Research	15.27
Small to Medium Business	12.81
Start up	7.79
Large Company (Over 250 employees)	7.30
Public Sector Organisation	6.56
Voluntary sector or not-for-profit organisation	4.84

Table 10: Possible options to select in order to define the type of organisation that is requesting the data

Suggested Use	% of requests
Research	52.12
Business Use	37.46
Personal Use	24.52
Community Work	13.43
Other	8.09

Table 11: Options for suggested use of the data

Organisation type. When making a request, users could self-report the organisation they belonged to. Table 10 shows that the largest group issuing data requests were individuals, making up over 45% of all requests. The next largest groups were users requesting data for academic and research purposes (15%) or users representing small to medium businesses (13%).

Suggested use. Users were asked to specify how they would use the data that they requested. Table 11 shows the list of options users could choose from. Research was the most popular declared use of the data (52%) which, combined with the fact that only 15% of requests were declared to be made on behalf of research and academic institutions, suggests that data is used for non-academic research purposes. Taking into account the high proportion of requests made by individuals (45%), much fewer - 24% of the requests - were declared to be for personal use. This indicates that individuals may look for data for business use, which was the second most popular use option.

Request motivation. In Table 12 we present the list of motivations that users could choose from when requesting the data. The inability to find the required data is the most popular reason: this justification is given for more than 40% of the requests. It is not possible to determine if the data was available and users could not find it, or if it was indeed missing from the portal. However, this could be an indication that portals which offer search need to improve their search functionalities; this is supported by our previous work [25]. We analyse the ways users talk about data in their data request (in Section

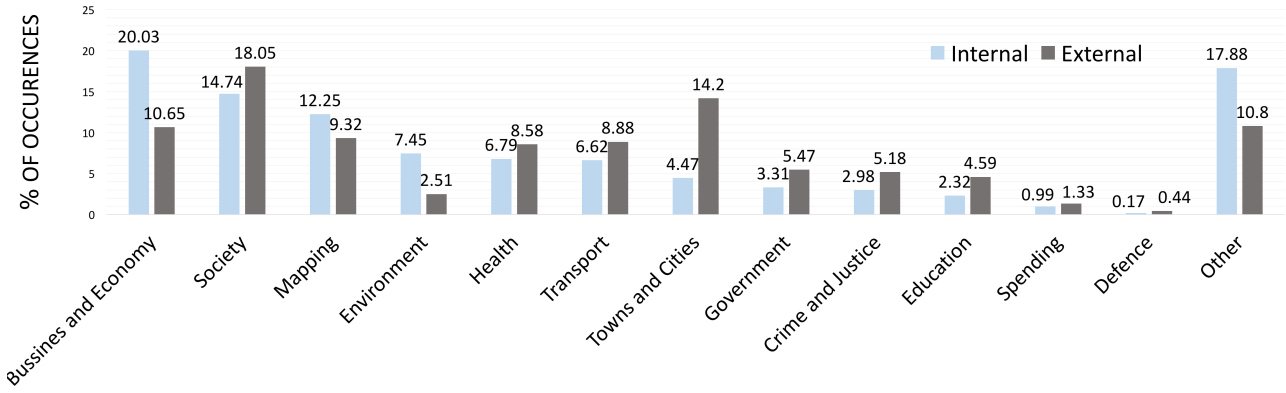

Figure 7: Distribution of topics within internal and external queries

Data Theme	%
Not able to find the data	41.75
The data is published but not in a format I can download and use (e.g. only displayed on-screen or only downloadable as a PDF rather than CSV)	8.15
The data is supposed to be published but the download links don't work	5.03
The data is not up-to-date	4.78
A version of the data is published but I need a different version	3.99
There are financial charges for the data	3.74
The data is available but the licensing terms are too restrictive	2.82
The data is subject to restrictions because of personal confidentiality	2.21
The data is subject to restrictions because of commercial confidentiality	1.59
Other	25.94

Table 12: List of reasons for making a data request with percentages of their popularity

6.1) to find commonalities pointing to relevant areas of improvement in such search functionalities. The second most popular reason given by users for issuing a data request was to request existing, published data in another format more suitable for their purposes. Other reasons for issuing data requests (e.g. financial charges for the data, broken links, or restrictive licences) were less frequent.

Time of issuing the request. Each data request contained a time stamp of the date and the exact time the data request was issued. We saw that majority of requests are done on weekdays (more than three times in comparison to weekend days). We also noticed that the majority of requests were issued between 9*am* and 6*pm*. This supports the results from user activity in Section 4, however, both groups represent different samples of users.

The thematic analysis resulted in two main categories describing different approaches of understanding the data requests. In Section 6.1 we describe the different attributes that were prevalent in the requests in order to present a comprehensive picture of the request content. For example geospatial or temporal information, format, license, etc. In Section 6.2 we present an analysis of the request context, in which people express content beyond the actual data they are looking for. This included for instance representation of motivation, comparisons, references to other datasets, analysis or specific questions they want an answer to.

6.1. Data Attributes

In this section we examine four data attributes that emerged as prominent themes from the data requests: geospatial information, temporal information, restrictions and information about granularity.

Geospatial information. (n = 77.5%) of requests included some reference to geospatial information at varied levels of detail and scope. They were asking either for information about several nations or larger areas such as the whole of the UK. In contrast to that many requests included specific points of location, such as a borough, a street or even a specific address, as can be seen below.

Q1: Would you provide me with any groundwater level data you have for the Preston area centred on grid reference SD 546 291?

Q2: Number of yearly conviction for all computer misuse offences in England and Wales for each year

since 2006, as defined under the Computer Misuse Act 1990.

Some users request location in specific granularity, as can be seen in Q3:

Q3: Could I request a full dataset of all the UK speed limits per road.

Location is expressed differently by different users and by their respective information need. We found geospatial information referred to as country or city names, ISO codes, abbreviations, latitude/longitude, grid references or other specialised identifiers. On the other hand people also used very vague terms, such as "overseas", "near to" or "surrounding area of". Some users do not seem to know what data is available and therefore try to narrow the scope of the location rather than specifying an exact location. This range of behaviours can be seen in quotes Q4 and Q5:

Q4: I am looking for shapefiles on general environmental data near Ferndale (Wales)

Q5: I would like to have access to all the data available up to nowadays regarding fish (where as it is monitoring, surveys, or any other type of data collected regarding fish) for Beane river downstream (lat 51.806014; long -0.066997) upstream (lat 51.981001; long -0.094448)

People often expressed geospatial needs or requirements by defining the boundaries of the area they are interested in (e.g. "London and surrounding areas"; "from Richmond to west Thurrock". They do this either by defining an area if a "name" is known, or by expressing borders between which their area of interest lies in.

Expressions of geospatial information in the requests are complex and show a large variety. This is partially due to the fact that there is no standard way of expressing geospatial information in natural language. There are also many domain specific geospatial boundaries in use which in addition have changed over time (such as currently unused historical boundaries). The way that we record administrative boundaries has changed and to make historic comparisons people search for a comparable area (e.g. Q6).

Some mentions of locations were focused on locations that are not directly understood as geographical areas. This can be seen in Q7, where a user is requesting information about specific zones.

Q6: I am looking at British voting patterns across the past three General Elections of 2005, 2010 and 2015 and comparing the vote shares of the key political parties [..] the shape of parliamentary constituencies changed from '05 to '10. I need both maps to sort data from one to the other. At the moment, I am unable to find the data for the previous, 2005, voting parliamentary constituencies anywhere online.

Q7: I want to do some simple mapping of flood risk and I the require latest flood zone data (zone 2 and 3, and flood defences) to import into a GIS. If possible historic flood zone data from a few years ago (1-10 years ago) would be great to offer a comparison for analysis.

There is a range of geospatial information needs represented in the requests. From what we could infer from the requests users often aim to obtain data about specific locations to inform their decisions, to integrate the information in an analysis that is focused on one ore several areas, or integrate the data into an existing service or application:

Q8: I would like to investigate the relationship between weather conditions and occurrence of potholes. For this study, data on historic pothole occurrence is needed. The last 10 years of data for the occurrence of potholes Birmingham area is requested.

Q9: I have been interested in the split of EU voters by region, however I feel that a more useful statistic would be "by place of birth" as some people vote away from their home in the UK or overseas. is this split possible?

Q10: I am doing a ground investigation report on the jubilee line extension investigation carried out during 1990 and would find the lidar data helpful for my report.

Note the contrast between the high number of requests including geospatial information and the comparatively low number of keyword queries containing them. This suggests that text search boxes are not appropriate for searches that include geospatial parameters.

Temporal information. Looking at temporal information (which we defined as every mention that includes a reference to a unit of time) – which was represented in (n = 44%) of the requests – among which majority (87.5%) of statements refer to the time period covered by the data, meaning the temporal boundaries of interest. These were expressed either from a point in time to

the current date, or between two boundaries, or requesting the most up-to-date data on a topic. These were represented in different levels of granularity. Other temporal information referred for instance to temporal information in the dataset, such as the age of a person or the time of an event.

These requests illustrate statements from a point in time to the present:

Q11: The period 20/5/2016 until most recent.

Q12: Number of deaths in the last 20 years where cause of death has been certified as cancer. Counts are required by age of death.

These requests are an example of two exact boundaries specified in the requests:

Q13: I would like to see all comprehensive school terms and holiday dates from every council in England for 2015/2016 and 2016/2017

Q14: I am requesting the microdata of this survey from year 2012 to 2014 (smoking and drinking habits). I am working in a research looking at smoking prevalence by birth cohort, age and year. I am using the General Lifestyle Survey from 2000-2011 and would like to continue the sequence until 2014

The following examples show temporal information, requesting the most up-to-date time period in which the data was recorded:

Q15: Most up to date Stop and Search data in the UK. Including Ethnicity.

Q16: I am interested in obtaining a complete up to date list of every licensed taxi and private hire operator in England, Scotland and Wales

Other statements containing temporal information specified the required granularity of the data, such as *daily/monthly/yearly* as can be seen in Q17:

Q17: Daily Average temperature UK 2014 to 2016

Q18: Inflation, from january/2006 till march/2013 (as monthly)

Q19: The time of the crime (to the second ideally, but just as accurate as we can get)

Others expressed the required time in vague terms:

Q20: Most recent and historical commercial property rents, by postcode, census output area or ward.

Some requests also mention temporal information in order to answer a question about a specific point in time. In that case temporal information from within the data is required to answer the query:

Q21: For which tax year was there the greatest inheritance tax revenue per head of population, in real terms? In this tax year, what were (a) the inheritance tax rates, and (b) the other major differences from this year's rules?

Q22: Amount of deaths in the last three years of those with learning disabilities.

The following example shows required temporal information from within the dataset when it refers to "children under 18 years":

Q23: Up to date statistics concerning children (under 18 years) smoking, drinking and taking drugs, attendance/ exclusion from school, and anti social behaviour statistics in the Hertfordshire area

Or the following example in which people asked for the dates of all bank holidays:

Q24: Data on all UK bank holidays, past and future. [..] Data should go as far back as reasonably practical - 1970 as a minimum, post war is desirable. Into the future, the data is obviously a prediction, but should cover 40 years in advance generated using current known rules for bank holidays.

In summary - temporal statements were used in several ways to define boundaries for the requested data: either to ask for the most current data; or from a point in the past to the present; or between two specific dates or for a certain number of years. These were presented at different levels of granularity as some statements specified a certain date and others were more vague mentioning e.g. "historic data". Another type of temporal information was represented when users were trying to answer specific temporal questions.

Restriction. To get data that will be useful for a specific task, users specify various constraints or restrictions on the requested data (n = 26.5%). Requests include statements specifying restrictions on the format of the data; price; specific data types; licence or a subset of data when a file was too big to use.

Below is an example of users specifying expectations about the price and license of the requested data:

Q25: There are licensees for the complete dataset, and queries from the dataset are of a suitable form, but quite expensive (5p per query). The cost would need to be in the fractions of pence or free to make this a viable usage.

Q26: Unique property reference number, and post code(s) are very important data assets (persistent identifiers) when it comes to empowering individuals to take more interest in their personal data, and data about them, and ultimately benefit from doing so. Both are locked up behind barriers created by history, and/ or failure to account for the impact of opening them up.

Q27: The licencing is very restrictive and does not allow for commercial use.

As Q28 and Q29 show it is crucial to publish data in a way that assures the possibility of it being useful for various tasks.

Q28: I guess Bank and IT Companies would be having. It would be helpful if not all the data at least the subset of original data is available.

Q29: Each NHS foundation trusts sends their annual financial data to Monitor in excel format. That data should be made available in excel format. Monitor currently publish a tiny subset in PDF format which is useless.

When file formats are mentioned they either relate to data being published in a non-machine readable format:

Q30: The data is published but not in a format I can download and use (e.g. only displayed on-screen or only downloadable as a PDF rather than CVS)

Or they specifies a format for the dataset that is needed for the respective task:

Q31: The data to be provided in a shape file format with the appropriate address(es) attached to the unique ID number

Granularity. We define granularity as the level of detail to which the data is broken down (e.g. data could be presented per kilometre, meter or centimetre). Requests often specified the desired level of granularity of the data (n = 24.5%). This was mostly found for temporal or geospatial granularity, but also subject granularity. For instance *"hourly weather and solar data set"*; *"25cm grid data"*; or *"prescription data per hospital"*. Below we present different ways granularity was expressed in the requests.

Users request data with specific granularity as this can be crucial to make the data actually useful for a specific task. For example in the case of Q32, if the data was presented by hospitals in specific boroughs or for all London hospitals it would fully miss the reason of the data requests – in which the crucial part was to have an overview of the data in a per hospital manner.

Q32: I would also like to know the number of accident and emergency admissions and births per hospital over a year.

Q33 is another example where the granularity of the data is highlighted as important, together with additional specifications of the lowest granularity required for the data to be suitable for the task.

Q33: I require a data set that shows the average daily temperature for the UK from 1 March 2014 to 31 July 2016. For example with the following columns: Date, Average Temperature (e.g. 01032014, 12). The data doesn't need to broken down any further than that.

Q34 and Q35 are examples of requests for data with the most detailed possible granularity - which could be particularly challenging for search functionalities to understand as they most often do not search within the dataset itself and granularity is challenging to express in metadata.

Q34: The research will identify whether there is a correlation (or not) between Road Traffic Accidents and Accidental Dwelling Fires and Youth Unemployment. The data that we are looking for should break down as much as possible e.g. into post codes.

Q35: A data source with sufficient accuracy to enable marker post references around the M25 to be located on the network. The data should be in the form of OSGRs or XY co-ordinates of sufficient accuracy

Lastly, another example of granularity information in data requests can be seen in Q36 where users requested data per day and want to be updated with new datasets on a monthly basis.

Q36: I want to find out how much it costs to run the London underground network each year. For example, how much does it cost to repair tracks each

year? How much does it cost to run each tube station, per day/week/month (preferably broken down into specific areas for example lighting, heating, maintenance, staff, etc.)

Q37: For our project we need to know the type of individual crimes as far back into the past as possible, and when they happened. Monthly crime records are not granular enough.

The level of granularity can be crucial for some the data to be useful. For example data that is aggregated to a country level would not be very useful for an analysis in a per city manner, although both datasets cover the same region. In current dataset search solutions defining the desired granularity of the dataset is not possible (unless it explicitly was stated in the description of the dataset).

6.2. Request Context

The analysis of the request context focuses on expressions framing the requirements associated with the information need, rather than the data required to fulfil it. This means statements expressing requirements or justifications, beyond the actual data that is requested. For example, this includes details such as users' motivation, comparisons or references to other datasets, and examples of specific questions users aim to answer with the data.

We grouped codes which were related to the expected structure of the data, and the type of expected outcome of the request (such as looking for a full dataset, looking for a particular data point, or looking for the results of an existing analysis). We also included mentions of specific headers, as well as pointers to other datasets that are similar to the requested one. In addition, we included requests that describe their rationale for seeking a particular type of data in more detail, as well as mentions of data quality. Below, we provide examples of each of these structures.

Representation and structure. Some requests contained detailed description of the dataset they are looking for (n = 32%) This was presented as a list of information –for instance a list of headers in the dataset. Others pointed to another dataset that presents information in a similar way as they are looking for; or pointed to a dataset that already exists but that still does not fulfil their information need (because of insufficient information, insufficient granularity or of different geospatial or temporal boundaries for their task).

In Q38 a user wanted to obtain the same dataset as one that was already published for a different time frame:

Q38: The river quality data that is available is limited to 2006. I currently need the same river quality data for the East Anglia region that is more recent, ideally as recent as possible (e.g. post-2010)

Q39 is an example of a specific list of information that is expected in the dataset:

Q39: For each of the schools under the Academy trust - the Head teacher, address, number of students, age range, telephone number.

Q40 shows a similar scenario, in which a user highlights a specific type of information that is missing in an existing dataset, but that is necessary for their task:

Q40: Accident Cause column in the data is missing, for example: Accident cause ="over speeding", "jumping a red light", "wrong overtaking", "lack of safe distance between vehicles"

In Q41 below we can see an example of a request for dataset that is similar to an existing dataset.

Q41: Details of all expenditure over £500 (or some other limit) on a monthly basis similar to the current publication of spend data by central government departments and local authorities.

When requesting a specific data structure, Q42 specifies data needed in a simple format that will not be challenging to analyse and understand.

Q42: I would like the overall cost of the UK Government published in simple format that the non accountancy literate among the electorate can understand. Ideally it would be set out as costs per themes as listed below BUT also show the complete cost to ensure that nothing is omitted.

The need for data in a specific structure or representation could be an indication of a need for implementing functionalities to search engines that support search for similar datasets to ones that get proposed by the user; or for supporting search over specific headers of the dataset. Further it could indicate the need for dataset recommendation systems that suggest datasets similar to ones already selected by the user, but fit their information need better. Requests in which the required headers or categories expected in a dataset are listed

indicates the usefulness of presenting the headers of a dataset to users in a search scenario.

Expected outcome. We further found that requests differ also in terms of their expected outcome. Some users expect specific data points or answers (n = 5.5%), others expect whole datasets (n = 78%) or the results of an analysis (n = 11.5%). This indicates the need for systems that provide a better support of datasets search functionalities. However, we those percentages can be biased by the nature of the data portals which provide uses with datasets only and does not support other functionalities (e.g. onsite data analysing tool or question answering functionalities) which could support a wide range of information tasks and a range of skill-sets amongst users.

Below we see an example of requests in which user express their information need as searching for already performed analysis instead of searching for a whole dataset:

> *Q44: I am currently investigating the number of hospitals, clinics, geriatric residencies pharmacies and laboratories across the UK and was wondering if a study could be done showing them per region and maybe a map of the UK, visually showing where they are gathered. Big circles on those regions with the most of them and smaller*

Q45 illustrates that users can expect an answer to a question which could be a single data point from an existing dataset (assuming that such a dataset exists):

> *Q45: Is there any statistics pertaining to the number or percentage of schools in the UK that are adhering to Prevent Duty in terms of IT/network security and firewall settings?*

Q46 and Q47 present requests for whole datasets:

> *Q46: All parking fines recorded by fine amount and location address of car parking*

> *Q47: listing of all the Academy Trusts with member schools of each trust (Primary and Secondary). For each trust the CEO (trust leader)/ address. For each of the schools under the Academy trust - the Head teacher, address, number of students, age range, telephone number.*

The way people express the desired outcome of their requests might be influenced by the semi-structured request form and the majority of entries are expressed in free-text. Search for data points is currently not supported on governmental open data portals.

Rationale. In 31% of the requests the underlying motivation was specified (e.g. *"I am a PhD student working on aquatic plants"*) or details of the analysis that is planned with the data (e.g. *"in order to show where (in Birmingham) there exists unemployment"*). In some requests, users specified that they want to compare the dataset they are requesting with one that they already have (e.g. to compare the income and expenses; compare spending to other London boroughs) and want to be supported in this process.

In Q48 we see personal reasons mentioned for the data request:

> *Q48: I am looking for a dataset available on all economic sanctions imposed by countries on each other for my master's dissertation. So I can run a regression analysis on the imposition of economic sanctions against rise or fall in GDP per capita of a country.*

Q49 and Q50 illustrate a description of planned analysis with the requested data:

> *Q49: I am trying to find map data for all local authorities in the UK (England, Wales, Scotland, NI) so I can render it on Google maps or OpenStreetMap.*

> *Q50: The research will identify whether there is a correlation (or not) between Road Traffic Accidents and Accidental Dwelling Fires and Youth Unemployment*

We hypothesise that for the majority of requests describing the rationale behind them, reasons are given due to the assumption that those requests will be read and assessed by people working at the data portal, as opposed to being screened automatically. This encourages users to describe their data needs in detail and in natural language. However, when indicated in a request that the data is needed for comparison (or to be combined) with two datasets, this may indicate value in implementing features to automatically assess the potential of combining two datasets (for instance based on the presence of the same header in each of the datasets).

Quality. Some requests indicated that a particular dataset has quality issues and they request the data in better quality. This included the one caused by a service providing the data (such as a broken link to a dataset)

and by the data itself. Quality can be understood on many different levels and is very dependent on the users' task [25]. Different users describe quality in different ways - fitting their information seeking scenario. Some requests mention quality of data indirectly by stating that a dataset has insufficient granularity or by requesting additional columns for a published dataset.

Below we present an overview of different mentions of quality in the data requests. However, the line between quality metrics and restrictions for or granularity of the data is not clear cut. Quality is mentioned mostly in relation to already existing data to criticise or explain why it is not useful for a particular information need, whereas restrictions or granularity are often expressed as requirements for requested data.

Data not being detailed enough (e.g. as in Q51) can be seen as one of the quality issues. Aggregation of data can result in it being not suitable for a task.

> *Q51: Met office publishes current data, but only historic averages, not historic data values. Without this history, I will have to wait years to amass sufficient current data for analysis.*

A similar issue for the usefulness of data is it's format (e.g. quote Q52). Many tasks require data to be of a specific format; for example geospatial data or a dataset saved as a PDF file.

> *Q52: The data set of the PCT boundaries they supplied, in KML format, has data quality issues and they no longer have access to their source of that data.*

Another group of quality issues are missing parts of datasets (e.g. Q53); or specific values of the dataset missing in an existing dataset (e.g. Q54 where the dates are missing); or in some cases errors within the existing data are mentioned (e.g. Q55 and Q56).

> *Q53: I require the IMD data which covers the North East. In this region there are several statistics missing from the 2011 IMD publications which related to the LSOA I was wondering why this data is missing and if you could provide a complete dataset.*

> *Q54: Accident Cause column is missing, for example: Accident cause ="over speeding", "jumping a red light", "wrong overtaking", "lack of safe distance between vehicles"*

> *Q55: There seems to be a gap in detailed LIDAR data available for the area where this golf club is based between hemel hempstead and St Albans. could this be updated?*

> *Q56: OSGR Eastings & Northings require to be 7/8 digit not 6 digit*

Quality awareness can be seen as understanding the state of the dataset; meaning if it is out-of-date or when the next update should happen, or if there are missing values and whether it is still usable for a certain type of analysis. This awareness allows users to judge the relevance of a dataset for an information need. Search functionalities could therefore allow users to judge certain aspects of data quality in the context of their task - potentially before downloading the dataset.

7. Discussion and Implications

Internal and External Queries. Our findings based on the analysis of user search behaviour when accessing data portals (described in Section 4) suggest that dataset search is a work-related activity. We found that most queries issued directly on the portals (i.e., the *internal* queries) were related to datasets in the area of *business* and *economy*. By contrast, external queries were topically more diverse, with topics such as *society* and *towns and cities* appearing regularly. We also noticed differences in the ratio of question queries - a larger percentage of external queries included question queries. This may indicate that different ways of accessing the portal could be related to different types of information needs (e.g., specific answers versus full datasets). Further analysis is needed to determine whether internal and external queries are indeed authored by distinct user groups and where these differences comes from. In our previous interview study [25] we found evidence that there is overlap between the two groups. From the point of view of description metadata, these results suggest that open data portals should focus on providing business related themes and concepts.

Our findings show that dataset search queries are generally short, on average one word shorter than web search queries, as per the 2011 report by Taghavi et al. [43]. We believe short queries potentially indicate that, currently, users do not expect that the search functionality will be able to provide relevant data for longer and more specific queries. It appears that users currently tend to treat the search box of a data portal as a starting point for further exploration. The categories and metadata attributes used in data portals as well as enabling linking between datasets or metadata properties could be key in improving dataset search functionalities.

We believe both temporal search, which is more prevalent than reported for general web search [34], and geospatial search [12] require better support. In both cases, relevant keywords can have different levels of granularity (e.g., months versus years, cities versus regions or countries), which is not always matched by the publishing practices of the data owners. While the data portals we analysed are location-bound to a country and most datasets hold national data, supporting question-answering and dataset search scenarios will require more advanced geospatial indexing and reasoning features. Queries including some indication of time were almost five times more frequent than in web search [34], suggesting that datasets have a stronger relationship to time than documents. DCAT already includes properties for temporal and geospatial description of datasets, and our findings suggests that providing fine-grained descriptions of these properties could improve search experience.

Data Requests. Data requests are issued by users with specific characteristics and are not directly comparable to the sample of data searchers represented in the log analysis. However, they complement our results by adding an in-depth perspective on information needs for data that are explained in more detail. We found that a large proportion of requests were issued by individuals, during weekdays and most were classified to be done for research purposes. However, only a small number of requests were issued by declared academics or researchers, which suggests that data is often used for non-academic research purposes potentially including private decision making contexts or business use. This somehow contradicts the fact that majority of portal users as described in Section 4 are using the portal in the work related environment. We recognised that the majority of data requests were issued during weekdays and within working hours. We hypothesise that people who issued data requests chose the *organisation type* on the requests form to be *individual* and not *business or academic/research institutions* as a way of not answering additional questions. This is because the form somehow suggested that declaring to be part of an organisation was not compulsory and might result in having to answer additional questions and so in more work. Another possibility is that people consider certain activities as research, even if they are done privately, *e.g.*, research about potholes to file a data-founded complaint to the council.The most common reason for issuing a request was specified as the inability to find suitable data, whereas less common reasons included, for instance, data that was available but in the wrong format or for the wrong time frame.

Over three quarters of all requests included Geospatial requirements. Geospatial requirements are specified through boundaries, which are expressed in varying levels of precision. As the requests were issued in natural language, this also included vague terms such as *overseas* or through use of more informal geographical definitions such as *tube zones*.

Almost half of all requests contained temporal information, the majority of which were requirements for a specific year, a particular time frame or simply for the most up-to-date data. These define the temporal boundaries for the information need. Temporal information is often discussed using non specific expressions in natural language and this is also reflected in the requests, such as *historic, or in the past*. Temporal information can also refer to specific attributes within the dataset, such as *diabetes people over 60 years*.

The high prevalence of temporal and geospatial information indicates the importance of these features for the fitness of use of data for a specific information need, supporting results of prior research [24] that identified importance of the time frame to which the data refers to. However, a new aspect that we identified is the relevance of the granularity of the geospatial or temporal information. Even if data that is topically relevant to an information need is available - if it has the wrong time frame or location - it becomes useless. In current dataset search solutions defining the desired granularity for the dataset is not possible (unless it was explicitly stated in the description of the dataset). Even if granularity is not provided as a facet in the search functionality, an overview of the available granularity of the data in the dataset could be presented in the metadata. Our findings show both the popularity of, and the complexity of expressing, geospatial or temporal boundaries for datasets, which suggests the need for designing more advanced search functionalities to cater for these attributes.

Other features prevalent in the requests concerned restrictions on the data. These refer for instance to the format or the size of the data, price restrictions or licence. Those can to some extent be resolved in providing functionalities on the data portal to change the data format, or by allowing users to select appropriate subsets of data. Those kind of issues could partially be resolved by providing both publishers and data users with additional information on the data publishing process (e.g. assigning appropriate licences to the data).

Data requests provided further insights into how people expect data that they are requesting to look like. This included specifying the headers that are expected in the dataset, or defining a certain format. This might be due

to limited technical skills or data literacy. Their task could also involve comparing the data to other datasets that they already have which is easier in a certain format or with comparable attributes. This suggests a number of potentially interesting directions for further research, such as recommendation systems for datasets based on similarity between datasets or that take a dataset as an input. This can also include the indexing of headers to make them discoverable for search functionalities, as well as the presentation of headers to users in a search scenario.

Our findings further point to several quality dimensions that are considered important in this context. This includes access to data, completeness and the amount of data available, characteristics that were covered in literature [36]. We believe that to judge the relevance of a dataset for a tasks users need to be aware of these characteristics. This suggests that the inclusion of basic quality dimensions in metadata or search result presentation could support the discovery process.

Dataset Search as a Vertical. In order for the data to be of use for an information need it must meet certain criteria. We believe that prior literature together with the findings of this work suggest that dataset search has unique characteristics which result in requirements that current dataset search functionalities do not fulfil. This suggests a large space for future research to improve current, and develop new, approaches for dataset search.

We believe that, in a retrieval scenario, datasets are complex to understand due to the ability to transform or analyse data, but also due to the different formats and structures that data can be stored in [25]. Key findings include the importance of boundaries in dataset search for different information types, especially geospatial and temporal information, as well as the granularity of available data. One aspect is that this information can be expressed in different ways, sometimes very specific and sometimes very vague – therefore, descriptions would need to index ranges as well as exact values, and search interfaces would need to be flexible enough to enable fuzzier queries. Some existing data portals already enable filtering datasets by geospatial coordinates inside a user-defined box. The U.S. national open data platforms includes the map preview of the geographical coverage of some of their datasets. UK's office for national statistics allows filtering time series data by custom periods of time, leveraging the fact that the underlying data is already in a time series format and it has a manageable size. However, further research is needed to extract the spatio-temporal characteristics of datasets for their addition to metadata descriptions, in particular for the case of Big datasets that might contain different entities and granularities

Many requests specified a data type and format, which underlines the complexity of data search in contrast to searching for documents. Majority of portals currently cover filtering through different dataset types, however it does not support known from the literature complex search activities. Many tasks with data involve comparing, contrasting or combining data with other data. This is reflected in the requests which often refer to other data, or specify that more than one dataset is needed for the task and that datasets need to be comparable (in terms of format, identifiers, etc.). Our findings suggest that often the successful retrieval of a dataset does not fulfil a users information need even if the retrieved dataset was fit for use, but can only be seen as a step towards it. Therefore functionalities supporting recommendations, or links to other datasets have the potential to be very valuable.

It is important to note that in terms of both geospatial and temporal information there could be more than one dataset equally relevant to a single information need. For example, when a user requests data from the "last 20 years", this could be returned as a number of equally relevant datasets, whether as one dataset covering the whole period, or as an individual dataset per year. Such requests could be fulfilled by automatically presenting an aggregation of the relevant datasets and, particularly if one of these datasets is not available, showing the timespan covered by the returned datasets.

8. Limitations of the study

Search Logs. Comparisons of different search log analyses present difficulties as concluded by [18] in their study comparing 9 search engines by their transaction logs. Even within web search, it is stated that findings resulting from the analysis of one search engine cannot be applied to all web search engines. Following this, any comparison of our results with web search needs to be seen with caution, due to the different nature of the collected data. However, we believe that including data from several countries and different audiences increases the generalisability of our analysis.

Our study is based on dataset search engines that are part of governmental open data portals. Further studies with other kinds of dataset search engines are required before drawing general conclusions. Query topics were annotated with the use of tags from one portal (data.gov.uk). We decided to use those categories as they present an overview of the content of governmental open data portals. Furthermore, as all portals used the

Google Analytics suite, we were subject to its session definition and identification algorithm.

As we did not have control on the analytics being collected by each data portal, we had different time frames and data for each one. In cases where all queries were considered, there is a bias towards DGU, as the portal with most available data. We had no means of detecting potential automated user agents which might have influenced some of our statistics. Finally, due to privacy considerations, we did not have access to a large number of external queries.

Data Requests. The objective of this analysis was to gather additional insights into how people ask for data when they are not constrained by the limitations of a current search environment. The set of data used in this study, the data requests, are real natural language articulations of people asking for data which is why we chose them for this study. However, they come with natural limitations. As these request span over five years we had no opportunity to follow up with people to understand what they meant. Some of the requests were relatively short and the topics were relatively domain and UK specific. The generalisability of the results is unclear, however we believe that these requests enable us to get unique insights into how people might articulate their information needs about data. Naturalistic information seeking tasks requiring data are not commonly reported in literature, which is why we believe the data requests are a valuable means to better understand how people search for data. It would be interesting in future work to compare these written requests, both in their structure as well as in their content, to requests for data in different digital environments. Finally, not much of consistent demographic information is collected alongside the requests. It is possible that the users making these requests represent a specific sample of the population.

9. Conclusion & Future Work

We presented an analysis of search log data for dataset retrieval, based on internal search logs of four national data portals, external search logs of two of the portals that were based in the United Kingdom, and data requests issued to one of those data portals, in order to understand how data portal users search for data and provide insight about what are the most important features of descriptive metadata from the point of view of data consumers. We analysed those three sets of data in order to answer the research questions (a) What are the characteristics of dataset queries? (b) How queries from data portals differ from general web search queries; (c) how people request data in a non-constrained form. Our findings can be summarised as: (i) Dataset queries are generally short. (ii) Dataset search seems to occur mostly in a work-related environment. (iii) There is a difference in topics, length and structure between dataset queries issued directly to data portals and dataset queries issued to web search engines. (iv) Data requests describe the data by using boundaries and restrictions about location, temporality, specific data type and/or specific granularity (v) Our analysis suggests that the prioritary properties to describe datasets are temporal and geospatial coverage, with varying levels of granularity. All of them already exist in current vocabularies. Our results suggest that efforts on automatic generation of dataset descriptions should be focused on these properties.

As future work, we would like to: (i) Automate the generation of spatio-temporal descriptions of datasets. (ii) Analyse query log data from commercial dataset search engines, to identify differences and similarities with this study. (iii) Extend our study to click-through data: knowing which dataset pages users visited after performing a search and if a user downloaded them can prove invaluable to evaluate the effectiveness of the dataset search. (iv) Create a dataset search corpus in order to evaluate dataset search engines.

Acknowledgement

This project is supported by the European Union Horizon 2020 program under the Marie Skłodowska-Curie grant agreement No. 642795. We would like to thank all the portals for providing us with data from their platforms.

References

[1] Agichtein, E., Brill, E., Dumais, S., Ragno, R., 2006. Learning user interaction models for predicting web search result preferences. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '06. ACM, New York, NY, USA, pp. 3–10.
URL http://doi.acm.org/10.1145/1148170.1148175

[2] Ai, Q., Dumais, S. T., Craswell, N., Liebling, D., 2017. Characterizing email search using large-scale behavioral logs and surveys. In: Proceedings of the 26th International Conference on World Wide Web. WWW '17. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 1511–1520.
URL https://doi.org/10.1145/3038912.3052615

[3] Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D. A., Frieder, O., 2004. Hourly analysis of a very large topically categorized web query log. In: SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and

Development in Information Retrieval, Sheffield, UK, July 25-29, 2004. pp. 321–328.
URL http://doi.acm.org/10.1145/1008992.1009048

[4] Bendersky, M., Croft, W. B., 2009. Analysis of long queries in a large scale search log. In: Proceedings of the 2009 Workshop on Web Search Click Data. WSCD '09. ACM, New York, NY, USA, pp. 8–14.
URL http://doi.acm.org/10.1145/1507509.1507511

[5] Broder, A., 2002. A taxonomy of web search. SIGIR Forum 36 (2), 3–10.
URL http://doi.acm.org/10.1145/792550.792552

[6] Bryman, A., 2006. Integrating quantitative and qualitative research: how is it done? Qualitative research 6 (1), 97–113.

[7] Cafarella, M. J., Halevy, A., Madhavan, J., 2011. Structured data on the web. Commun. ACM 54 (2), 72–79.
URL http://doi.acm.org/10.1145/1897816.1897839

[8] Cafarella, M. J., Halevy, A., Wang, D. Z., Wu, E., Zhang, Y., Aug. 2008. Webtables: Exploring the power of tables on the web. Proc. VLDB Endow. 1 (1), 538–549.
URL http://dx.doi.org/10.14778/1453856.1453916

[9] Clark, M., Kim, Y., Kruschwitz, U., Song, D., Albakour, D., Dignum, S., Beresi, U. C., Fasli, M., De Roeck, A., 2012. Automatically structuring domain knowledge from text: An overview of current research. Information Processing and Management 48 (3), 552–568.

[10] Corbitt, B. J., Thanasankit, T., Yi, H., 2003. Trust and e-commerce: a study of consumer perceptions. Electronic commerce research and applications 2 (3), 203–215.

[11] Ermilov, I., Ngomo, A.-C. N., 2016. TAIPAN: Automatic Property Mapping for Tabular Data. In: Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management. Vol. 10024.

[12] Gan, Q., Attenberg, J., Markowetz, A., Suel, T., 2008. Analysis of geographic queries in a search engine log. In: Proceedings of the First International Workshop on Location and the Web. LOCWEB '08. ACM, New York, NY, USA, pp. 49–56.
URL http://doi.acm.org/10.1145/1367798.1367806

[13] Gonzalez, H., Halevy, A. Y., Jensen, C. S., Langen, A., Madhavan, J., Shapley, R., Shen, W., Goldberg-Kidon, J., 2010. Google fusion tables: Web-centered data management and collaboration. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. SIGMOD '10. ACM, New York, NY, USA, pp. 1061–1066.
URL http://doi.acm.org/10.1145/1807167.1807286

[14] Guha, R. V., Brickley, D., Macbeth, S., Jan. 2016. Schema.org: Evolution of structured data on the web. Communications of the ACM 59 (2), 44–51.
URL http://doi.acm.org/10.1145/2844544

[15] Guy, I., Ur, S., Ronen, I., Weber, S., Oral, T., 2012. Best faces forward: a large-scale study of people search in the enterprise. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM.

[16] Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A., Decker, S., 2011. Searching and browsing linked data with swse: The semantic web search engine. Web semantics: science, services and agents on the world wide web 9 (4), 365–401.

[17] Jansen, B. J., Spink, A., 2005. An analysis of web searching by european alltheweb.com users. Information Processing and Management 41 (2), 361–381.
URL http://dx.doi.org/10.1016/S0306-4573(03)00067-0

[18] Jansen, B. J., Spink, A., 2006. How are we searching the world wide web?: A comparison of nine search engine transaction logs. Information Processing and Management 42 (1), 248–263.
URL http://dx.doi.org/10.1016/j.ipm.2004.10.007

[19] Jansen, B. J., Spink, A., Saracevic, T., 2000. Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing and Management 36 (2), 207–227.
URL http://dx.doi.org/10.1016/S0306-4573(99)00056-4

[20] Jiang, D., Pei, J., Li, H., 2013. Mining search and browse logs for web search: A survey. ACM Transactions on Intelligent Systems and Technology 4 (4), 57:1–57:37.
URL http://doi.acm.org/10.1145/2508037.2508038

[21] Kacprzak, E., Koesten, L. M., Ibáñez, L.-D., Simperl, E., Tennison, J., 2017. A Query Log Analysis of Dataset Search. Springer International Publishing, Cham, pp. 429–436.
URL https://doi.org/10.1007/978-3-319-60131-1_29

[22] Kaur, N., Aggarwal, H., Mar 2018. Query based approach for referrer field analysis of log data using web mining techniques for ontology improvement. International Journal of Information Technology 10 (1), 99–110.
URL https://doi.org/10.1007/s41870-017-0063-2

[23] Kelly, D., 2009. Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval 3 (1-2), 1–224.
URL http://dx.doi.org/10.1561/1500000012

[24] Kern, D., Mathiak, B., 2015. Are there any differences in data set retrieval compared to well-known literature retrieval? In: Research and Advanced Technology for Digital Libraries - 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015. Proceedings. pp. 197–208.
URL https://doi.org/10.1007/978-3-319-24592-8_15

[25] Koesten, L. M., Kacprzak, E., Jenifer, T., Simperl, E., 2017. The trials and tribulations of working with structured data - a study on information seeking behaviour. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. CHI '17. ACM, New York, NY, USA.

[26] Kunze, S. R., Auer, S., sep 2013. Dataset Retrieval. In: 2013 IEEE Seventh International Conference on Semantic Computing.
URL http://ieeexplore.ieee.org/document/6693487

[27] Kwok, C. C. T., Etzioni, O., Weld, D. S., 2001. Scaling question answering to the web. ACM Transactions on Information Systems 19 (3), 242–262.
URL http://doi.acm.org/10.1145/502115.502117

[28] Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., Kruschwitz, N., 2011. Big Data, Analytics and the Path From Insights to Value. MIT Sloan Management Review 52 (2).

[29] Lehmberg, O., Ritze, D., Meusel, R., Bizer, C., 2016. A large public corpus of web tables containing time and context metadata. In: Proceedings of the 25th International Conference Companion on World Wide Web.
URL http://www2016.net/proceedings/companion/p75.pdf

[30] Li, X., Liu, B., Yu, P. S., 2010. Time Sensitive Ranking with Application to Publication Search. Springer, New York, pp. 187–209.
URL https://doi.org/10.1007/978-1-4419-6515-8_7

[31] Mitlöhner, J., Neumaier, S., Umbrich, J., Polleres, A., 2016. Characteristics of open data CSV files. In: 2nd International Conference on Open and Big Data, OBD 2016, Vienna, Austria, August 22-24, 2016. pp. 72–79.
URL https://doi.org/10.1109/OBD.2016.18

[32] Narang, K., Dumais, S. T., Craswell, N., Liebling, D., Ai, Q., 2017. Large-scale analysis of email search and organizational

strategies. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. CHIIR '17. ACM, New York, NY, USA, pp. 215–223.
URL http://doi.acm.org/10.1145/3020165.3020175

[33] Neumaier, S., Umbrich, J., Polleres, A., 2016. Li ing data portals to the web of data.

[34] Nunes, S., Ribeiro, C., David, G., 2008. Use of temporal expressions in web search. In: European Conference on Information Retrieval. Springer, pp. 580–584.

[35] Ortiz-Cordova, A., Yang, Y., Jansen, B. J., ???? External to internal search: Associating searching on search engines with searching on sites.

[36] Pipino, L. L., Lee, Y. W., Wang, R. Y., Apr. 2002. Data quality assessment. Commun. ACM 45 (4), 211–218.
URL http://doi.acm.org/10.1145/505248.506010

[37] Robson, C., McCartan, K., 2016. Real world research. John Wiley & Sons.

[38] Sebastiani, F., 2002. Machine learning in automated text categorization. ACM computing surveys (CSUR) 34 (1), 1–47.

[39] Sieg, A., Mobasher, B., Burke, R., 2007. Web search personalization with ontological user profiles. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. CIKM '07. ACM, New York, NY, USA, pp. 525–534.
URL http://doi.acm.org/10.1145/1321440.1321515

[40] Silverstein, C., Marais, H., Henzinger, M., Moricz, M., 1999. Analysis of a very large web search engine query log. In: ACm SIGIR Forum. Vol. 33. ACM.

[41] Spink, A., Ozmutlu, S., Ozmutlu, H. C., Jansen, B. J., 2002. U.s. versus european web searching trends. In: ACM Sigir Forum. Vol. 36. ACM.

[42] Spink, A., Wolfram, D., Jansen, M. B., Saracevic, T., 2001. Searching the web: The public and their queries. Journal of the American society for information science and technology 52 (3).

[43] Taghavi, M., Patel, A., Schmidt, N., Wills, C., Tew, Y., 2012. An analysis of web proxy logs with query distribution pattern approach for search engines. Computer Standards & Interfaces 34 (1).

[44] Thomas, D. R., 2006. A general inductive approach for analyzing qualitative evaluation data. American journal of evaluation 27 (2), 237–246.

[45] Vandic, D., Frasincar, F., Kaymak, U., 2013. Facet selection algorithms for web product search. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, pp. 2327–2332.

[46] Vandic, D., van Dam, J.-W., Frasincar, F., 2012. Faceted product search powered by the semantic web. Decision Support Systems 53 (3), 425 – 437.
URL http://www.sciencedirect.com/science/article/pii/S0167923612000681

[47] Verhulst, S., Young, A., 2016. Open data impact when demand and supply meet. Tech. Rep. March, GOVLAB.

[48] Weerkamp, W., Berendsen, R., Kovachev, B., Meij, E., Balog, K., de Rijke, M., 2011. People searching for people: Analysis of a people search engine log. SIGIR '11. New York, NY, USA.
URL http://doi.acm.org/10.1145/2009916.2009927

[49] White, R. W., Wang, S., Pant, A., Harpaz, R., Shukla, P., Sun, W., DuMouchel, W., Horvitz, E., 2016. Early identification of adverse drug reactions from search log data. Journal of Biomedical Informatics 59, 42 – 48.
URL http://www.sciencedirect.com/science/article/pii/S1532046415002427

[50] Yu, P. S., Li, X., Liu, B., 2005. Adding the temporal dimension to search " a case study in publication search. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. WI '05. IEEE Computer Society, Washington, DC, USA, pp. 543–549.
URL http://dx.doi.org/10.1109/WI.2005.21

[51] Zhang, W., Yoshida, T., Tang, X., 2008. Tfidf, LSI and multi-word in information retrieval and text categorization. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Singapore, 12-15 October 2008. pp. 108–113.
URL https://doi.org/10.1109/ICSMC.2008.4811259